Introduction au Deep Learning
2018-2019

ISAE, Toulouse
3ème année, Filière sciences de la décision

Logique du cours

Introduction au Deep Learning:
« Vue de dessus »

Réseaux de Neurones Artificiels,
Deep learning et CNN, Keras

Exemples de “Buzz” autour du Deep Learning

2011: reconnaissance de panneaux de signalisation : DL meilleur que l’oeil humain

2012:

WIRED Google’s Artificial Brain Learns to Find Cat Videos

2016: Alpha Go

© reuters/ Kim Hong Ji

© Telerama

MIT Technology Review

Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

A technique called deep learning could help Facebook understan its users and their data better.

2016: Y. Le Cun at The College de France

Mais que s’est-il donc passé pour que le Deep Learning soit « à la mode »?

Quelques dates-clefs de l'IA pour la partie learning: 1956->199x

- Motivation initiale « neuromimétique »: le cerveau effectue sans effort apparent des tâches complexes pour les ordinateurs.

- 1956: Darmouth Workshop et terme Intelligence Artificielle
- 1958 : **Perceptron** de Rosenblatt.
- *1973: Premier hiver de l'IA (Lighthill report)*

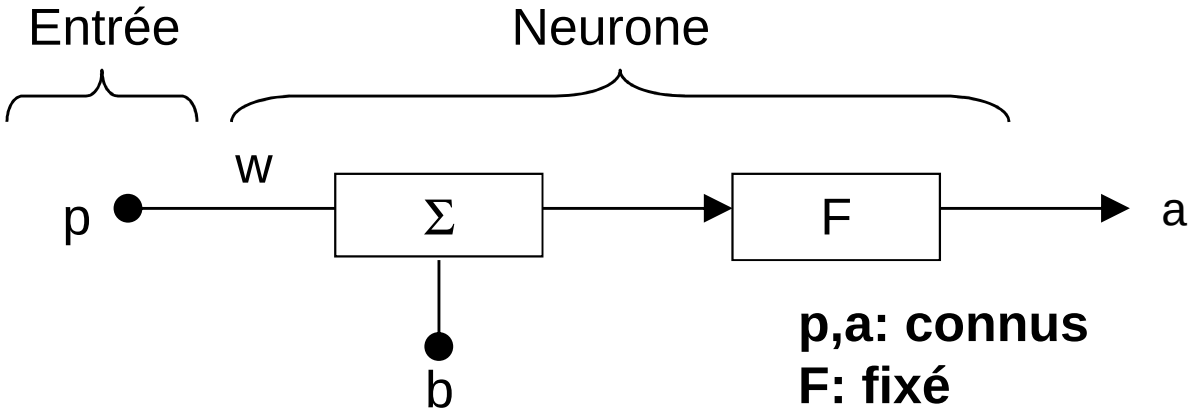
- 1986 : Perceptron Multi-couches de Rumelhart.
  Multi-Layers Perceptron (**MLP**), Rétropropagation (**Backpropagation**).
- *199x: Second hiver de l'IA*

Bref rappel sur les Réseaux de Neurones (jusqu'aux Années 199x)

Pour plus de détails voir Cours E. Rachelson

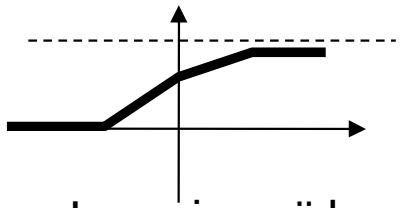
Rappels RN MLP (1/4)

- Le neurone formel à une seule entrée x est une fonction non-linéaire F, appelée fonction d'activation, de x pondérée par le poids w (« synaptique ») et le biais b.



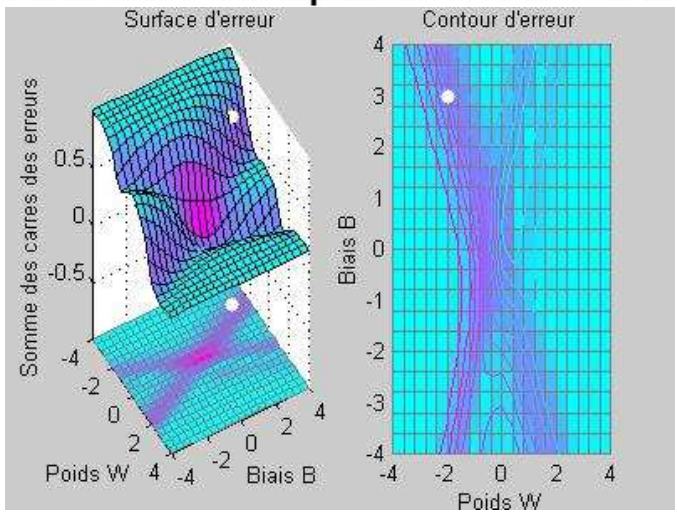
p,a: connus
F: fixé
W,b à estimer

Exemple de F :



Log-sigmoïde :
1/(1+exp(-W*p + b))

- En supervisé: Le poids w et le biais b sont ajustables en fonction du problème. pour que le réseau ait un certain comportement vis à vis d'entrées fixées.

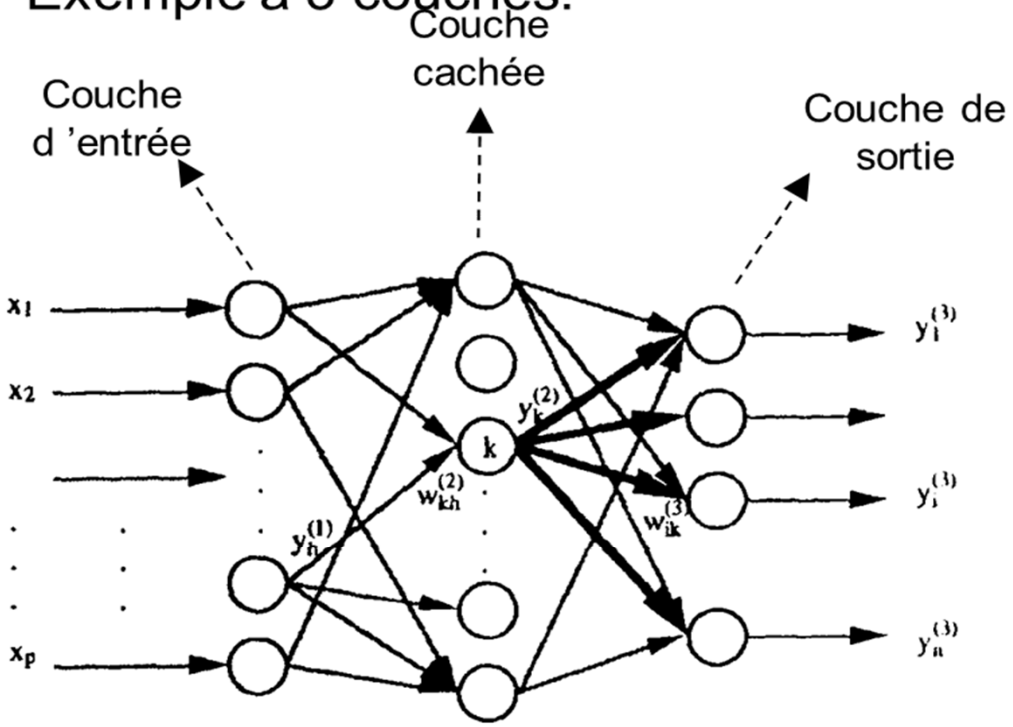


Système surdimensionné:
Plus d'équations (couples p,a) que de variables à estimer (w,b)
Estimation W,b par minisation critère .Exemple :
Critère moindres carrés par méthode du gradient.

Rappels RN MLP (2/4)

- L'intérêt des neurones résides dans les propriétés qui résultent de leur association en réseaux par couches, c'est-à-dire de la *composition* des fonctions non linéaires réalisées par chacun des neurones : **MULTI-LAYERS PERCEPTRON (MLP)**

- Exemple à 3 couches:



Estimation des paramètres du RN:

On connaît l'erreur quadratique (estimée-connue)^2 de la dernière couche => on sait estimer le gradient pour cette dernière couche.

On ne connaît pas directement l'erreur quadratique associée à chaque couche cachée (i.e. interne):

Comment faire ?

Rappels RN MLP (3/4)

- On utilise la **rétropropagation du gradient:**

- On montre que l 'on sait calculer l 'erreur associée à un neurone k d 'une couche quelconque j à partir des erreurs de la couche suivante j+1 => On évaluera le gradient de la fonction de coût en partant de la dernière couche vers la première.

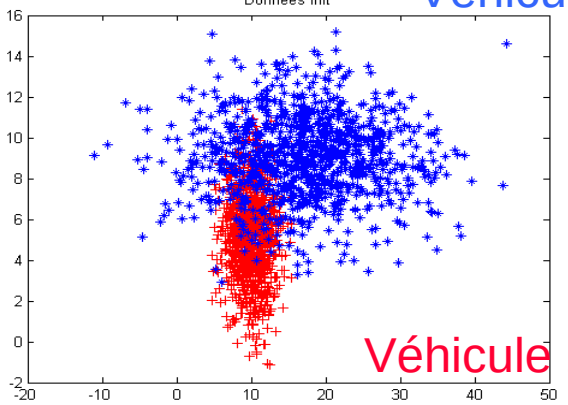
- Une fois ce gradient ∇ calculé, on peut mettre à jour les poids, par exemple:
  - Par la méthode du gradient simple : $w(i) = w(i-1) - \mu i . \nabla J(w(i-1))$ avec μi = pas du gradient
  - Par les méthodes de gradient de second ordre.

Rappels RN MLP (4/4)

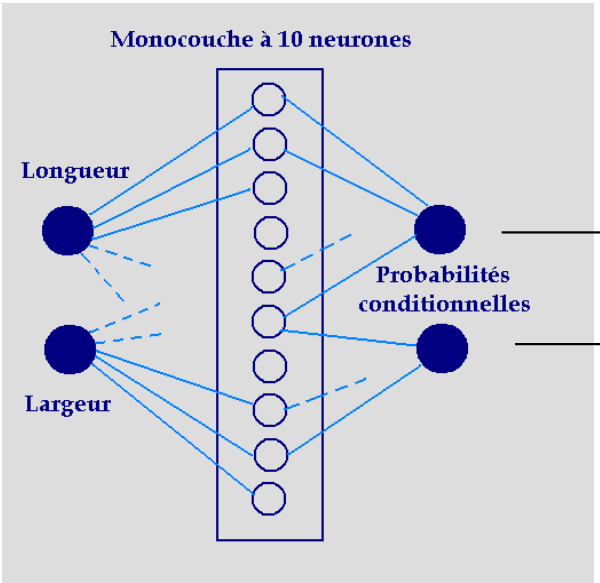
Image

Extraction Longueur et Largeur du Véhicule (**Features**)

Traitement d'images



Véhicule 1
ou
Véhicule 2



Ou en est-on en 199x ?

- Réseaux de Neurones (RN) :
  - ➔On est dans les méthodes « **Data driven** » différentes des « Model-based ».
  - ➔On sait estimer des architectures type MLP, Réseaux récurrents, Self-organizing maps de Kohonen etc. pour des problèmes de classification, segmentation, etc.
  - ➔Performances acceptables mais ne transcendent pas la communauté scientifique
  - ➔Les RN sont souvent présentés comme des « boites noires » => mauvaise réputation

- Autres aspects de l'IA
  - ➔Systèmes experts (LISP), Knowledge-based, Prolog
  - ➔Model-based
  - ➔…

- L'IA a finalement mauvaise réputation et on l'appelle même par d'autres noms. *199x: Second hiver de l'IA*

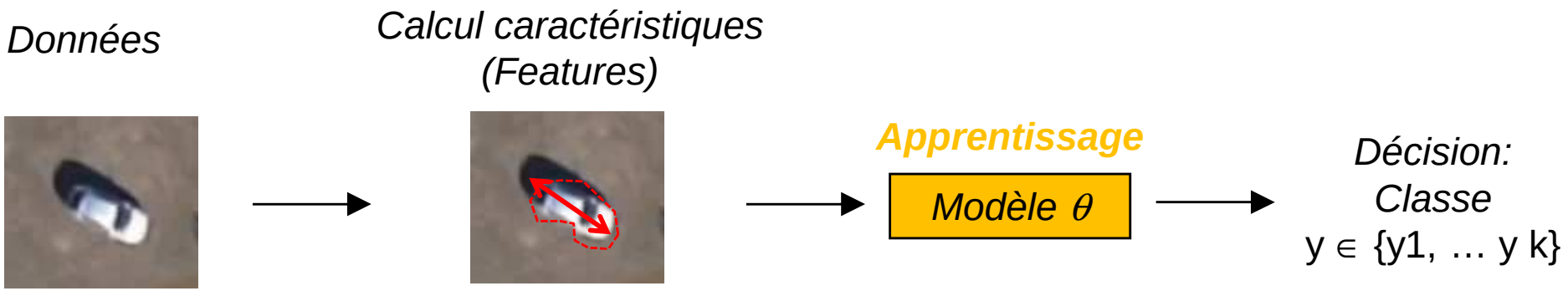
Mais des travaux d'irréductibles continuent

- 1981: Fukushima Neocognitron
- 1988: Le Deep learning avec les Convolutional Neural Network (**CNN**) de Y. LeCun

- 1997 : Deep Blue vs Kasparov

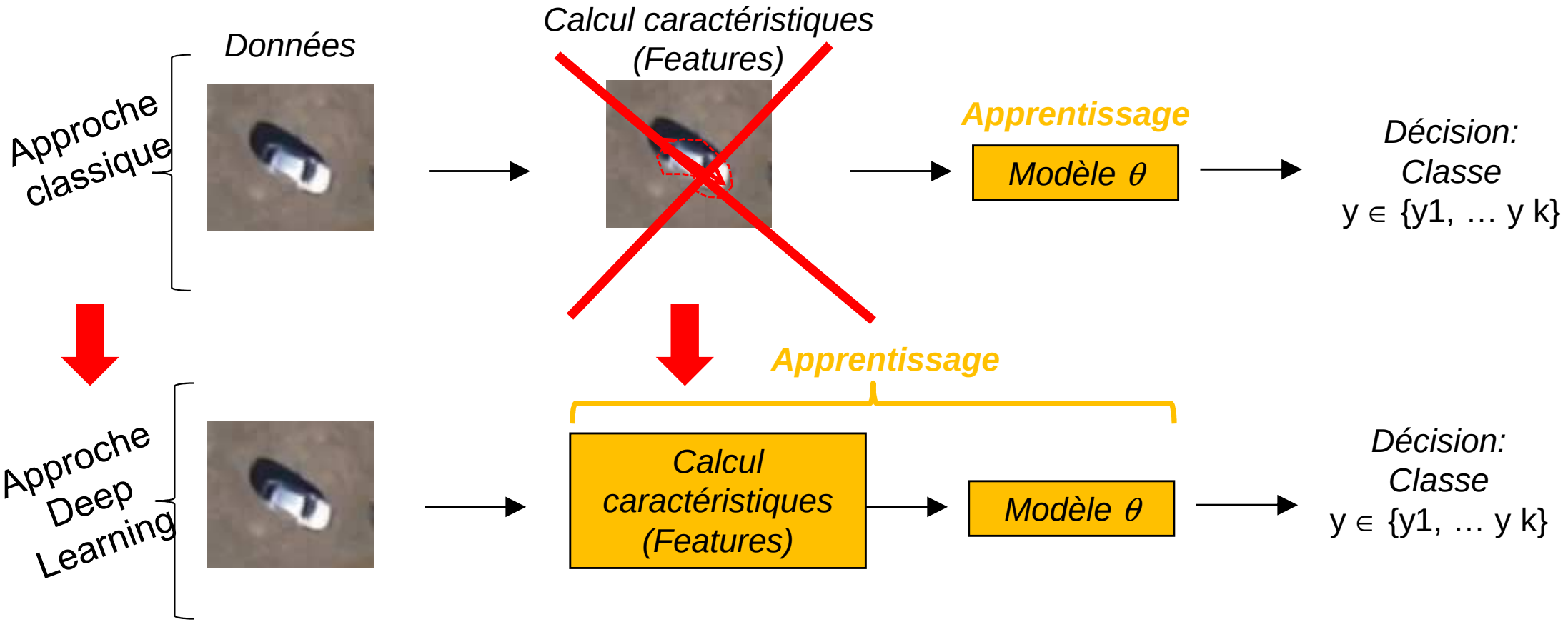
Idée générale du Deep Learning avec les CNN (Y. Le Cun 1988)

Machine Learning supervisé “classique”



- On passe beaucoup d’énergie sur la représentation des données (features): Quel vecteur est caractéristique des données pour prendre une décision ?
- La chaîne de traitement résultante est très spécifique au problème => A chaque nouveau problème on recommence.
- La performance/robustesse n’est pas toujours au rendez-vous.

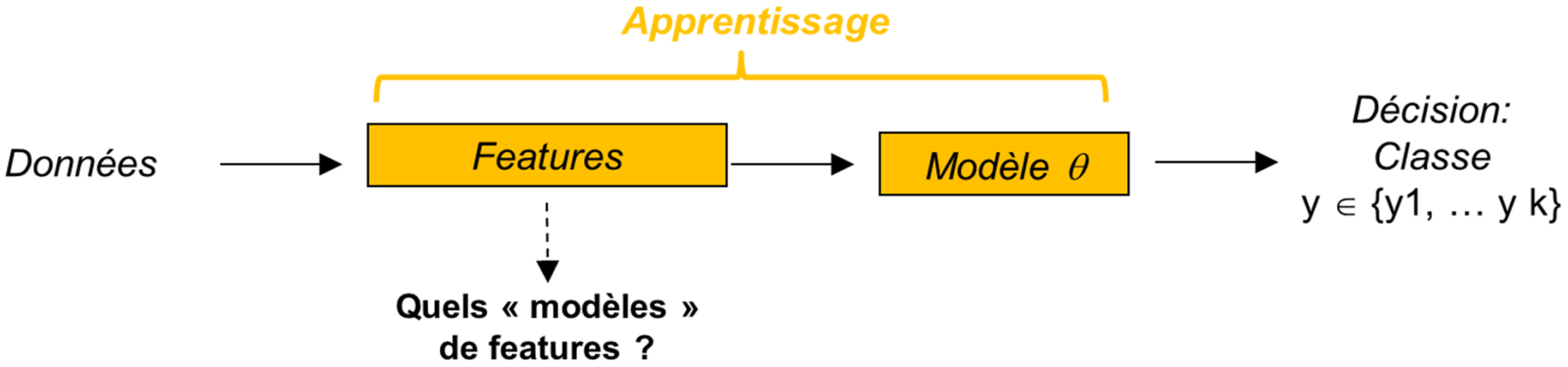
Que propose le Deep Learning avec les CNN ?



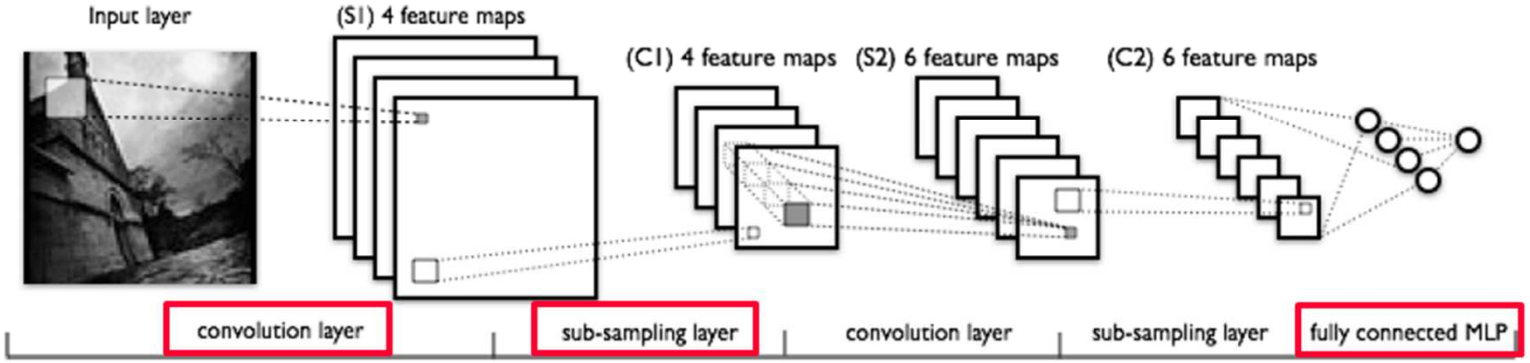
- On veut apprendre conjointement le modèle θ et la représentation des données (features) => la représentation des données fait partie des paramètres à estimer.

CNN: Travaux Y. Lecun

- **On veut donc apprendre conjointement le modèle θ et les Features**



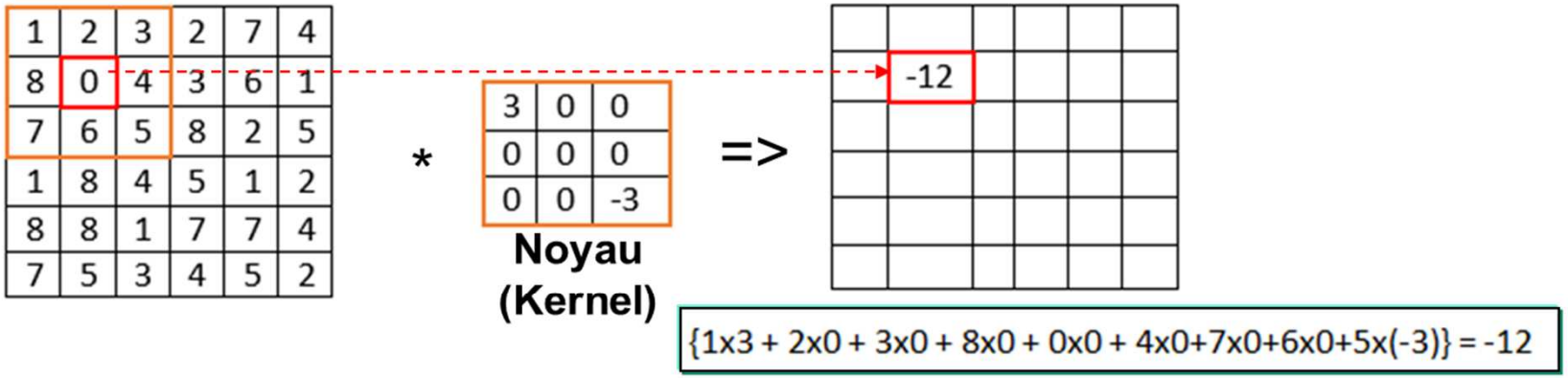
- **En 1988 Y. Le Cun & al , propose une architecture de RN basée sur plusieurs couches de traitement comprenant des convolutions et des sous-échantillonnages (Pooling).**



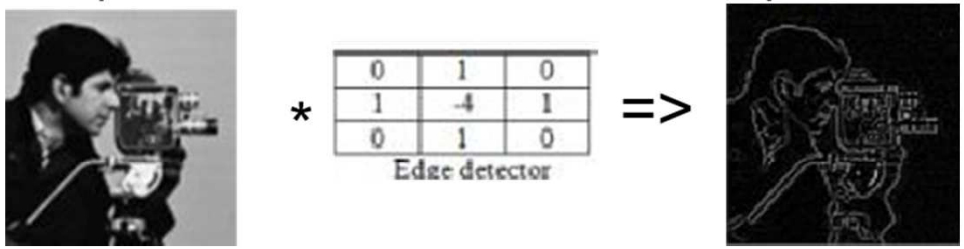
(Figure from http://deeplearning.net/tutorial/lenet.html)

Convolution

- **Pourquoi la convolution ?**
  - La convolution est un outil très utilisé en traitement d'image et plutôt« générique ».
  - L'idée est de changer l'intensité d'un pixel en fonction de celles des pixels voisins. Le genre d'opération effectué par une convolution en traitement d'images et le suivant:



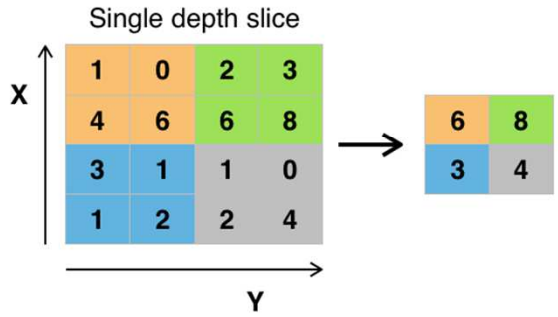
  - La convolution et permet de faire de nombreuses opérations : contours (edges), etc.



  - En un certain sens c'est donc un outil assez générique pour estimer des features et c'est une des raisons pour laquelle elle a été sélectionnée dans les couches des CNN.

Pooling

- **Pourquoi le sous-échantillonnage (Pooling)?**
  - Exemple de sous-échantillonnage par max pooling

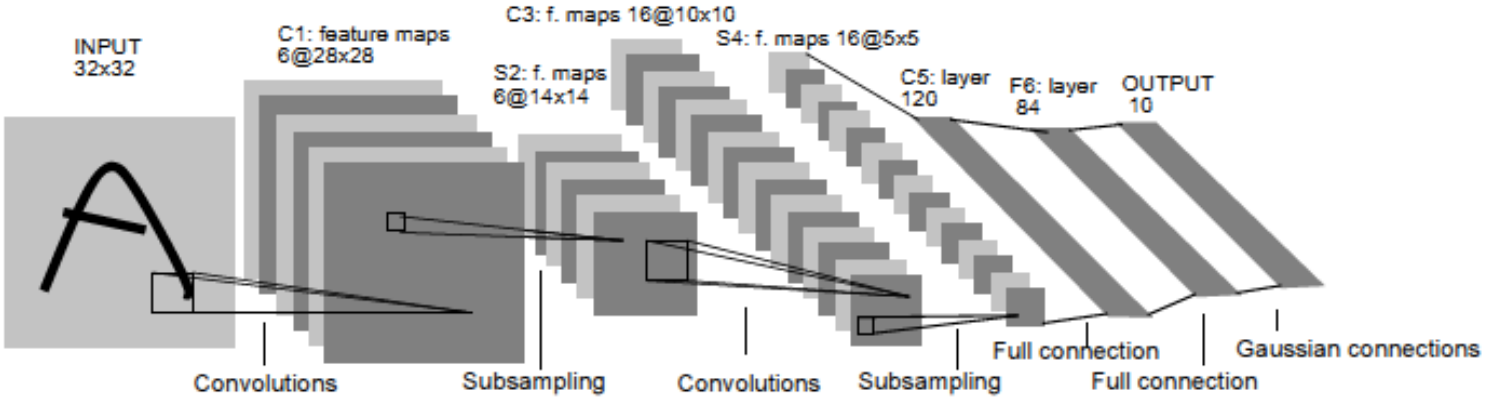


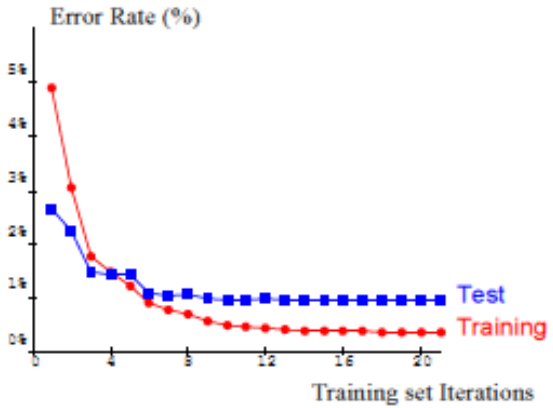
  - Réduire la représentation des données permet de réduire la complexité du modèle à estimer et aussi de réduire la charge de calcul et contrôler l'overfitting.
  - Intuitivement on accorde plus d'importance à la position relative des features qu'à leur localisation exacte.
  - Cela apporte un certain niveau d'invariance à la translation.

- **Quel Pooling ?**
  - Un article intéressant fait une analyse théorique sur les différents type de pooling et montre en particulier l'intérêt du max pooling:
    « *A theoretical analysis of features pooling for visual recognition* ».

Trait. Images MNIST

- **Exemple sur Base d'apprentissage/test MNIST :**
  - Caractères écrits
  - 60000 exemples en learning et 10000 en test.





Gradient based learning applied to document recognition Y .Le Cun IEEE nov. 1998

Mais pourquoi la communauté scientifique ne s'est pas jetée sur ces techniques en 1988 ?

- Comme on le disait
  - ➔Pour les RN types MLP Performances acceptables mais ne transcendent pas la communauté scientifique
  - ➔Les RN sont souvent présentés comme des « boites noires » => mauvaise réputation
  - ➔L'IA a finalement mauvaise réputation

- De plus
  - ➔Ces algorithmes sont gourmands en terme de données d'apprentissage
  - ➔Les calculs pour estimer les modèles sont coûteux

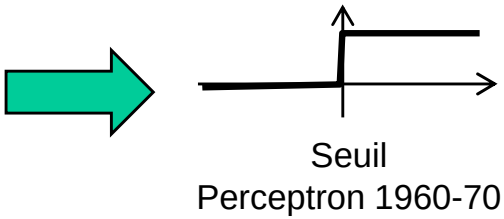
Mais alors que s'est-il donc passé?

Mais que s'est-il donc passé pour transformer l'essai?

- **Hors RN**:
  - Arrivée d'internet (199x) et les géants du numérique: Gafa, etc
  - Accès à des quantités de données gigantesques (« Big data »).
  - Accès à des puissances de calcul phénoménales: Cloud, HPC, etc.

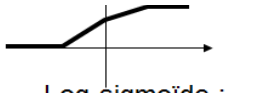
- **Concernant les RN**:
  - Modèles et Apprentissage/Généralisation plus évolués. Maîtrise accrue archi/régularisation:
    - Vanishing gradient

Evolution fonctions activation



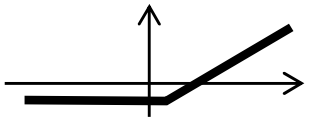
Seuil
Perceptron 1960-70

MLP 1986



Log-sigmoïde :
1/(1+exp(-W*p + b))

Rectified Linear Unit (ReLu) Lu : f(x) =max(0,x)
Deep Learning



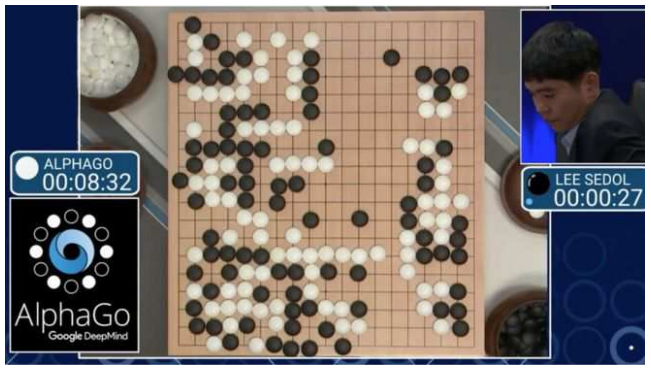
    - Eviter l'overfitting: Dropout, Regularisation, Data augmentation, etc.
  - Les DL montrent des résultats assez spectaculaires en traitement d'images

- **Communauté open source, buzz** :
  - La communauté scientifique du DL met à disposition de nombreux codes et supports didactiques.
  - Pas besoin d'être un spécialiste de l'image ou de la classification pour utiliser le DL
  - Effets médiatiques: Facebook, Google, etc.

Quelques dates-clefs de l'IA pour la partie learning: 1981->Aujourd'hui

- 2011: Watson (IBM) vs Jeopardy.
- 2011: Challenge Panneaux signalisation: Performances supérieures à l'humain
- 2012: Apprentissage non supervisé utilisant YouTube
- 2016: Alphago bat le champion du monde de Go.
- 2017: AlphaGo Zero bat AlphaGo par 100 jeux à zéro

2011 : Watson vs Jeopardy!



2016 :AlphaGo vs Lee Sedol

Ce que l’on disait en cours sur l’interprétation d’images il y a environ 15 ans: « c’est de la science fiction »

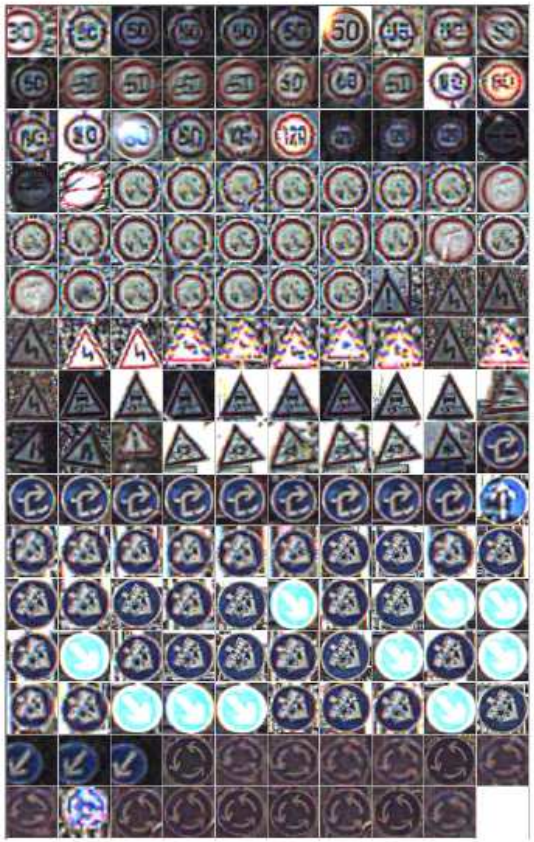
VIDEO
https://www.youtube.com/watch?v=keFdcSjcN3k

Un véhicule autonome?

VIDEO
https://www.youtube.com/watch?v=lowcgokiRG8

Trait. Images panneaux de signalisation

- **Exemple sur Base d'apprentissage/test Panneaux signalisation:**



Layer	Type	# maps & neurons	kernel
0	input	1 or 3 maps of 48x48 neurons	
1	convolutional	100 maps of 46x46 neurons	3x3
2	max pooling	100 maps of 23x23 neurons	2x2
3	convolutional	150 maps of 20x20 neurons	4x4
4	max pooling	150 maps of 10x10 neurons	2x2
5	convolutional	250 maps of 8x8 neurons	3x3
6	max pooling	250 maps of 4x4 neurons	2x2
7	fully connected	200 neurons	
8	fully connected	43 neurons	

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011). The German traffic sign recognition benchmark: A multi-class classification competition. In International joint conference on neural networks (pp. 1453–1460). IEEE Press

Exemple de codage d'un MLP sur MNIST avec Keras

```
model = Sequential()
# input: 28x28 images with 1 channels-> (1, 28, 28) tensors.
# this applies 32 convolution filters of size 3x3 each.
model.add(Convolution2D(32, 3,3, border_mode='valid',
input_shape=(28, 28,1)))
#model.add(Activation('relu'))
#model.add(Convolution2D(32, 3,3))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Convolution2D(32, 3,3))
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))

model.add(Flatten())
#model.add(Dense(128))
#model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(10))
model.add(Activation('softmax'))

sgd = SGD(lr=0.1, decay=1e-6, momentum=0.9, nesterov=True)
model.compile(loss='categorical_crossentropy',
optimizer=sgd,metrics=['accuracy'])

model.fit(trainimages, trainlabels, batch_size=32, nb_epoch=1)

score = model.evaluate(testimages, testlabels, batch_size=16)
```

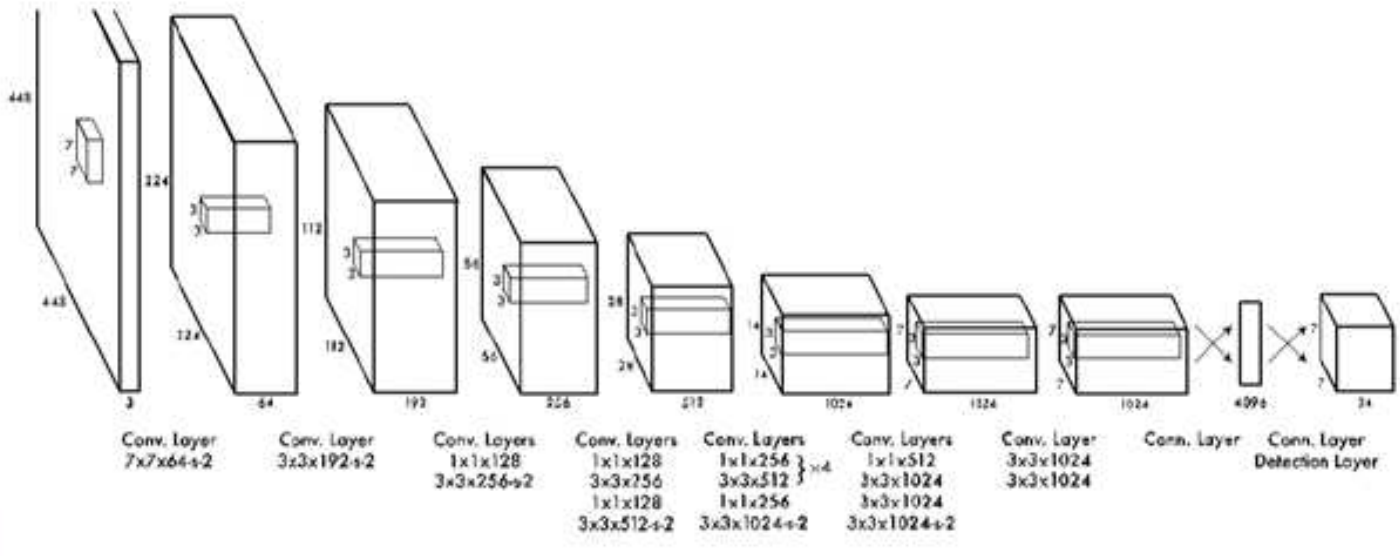
Un premier écueil à passer

« Ca ne marchera jamais car il faut des millions d'exemples avec des photos (*de chats par exemple*) mais on n'y arrivera pas sur des applications plus industrielles »

Et donc … ?

Une réponse: le Transfer Learning

- Utilisation de modèles pré-entrainés sur des grosses BD d'imagettes
  - Directement pour de la classification
  - Comme extracteur de features pour une classification avec RF, MLP, etc.



J'utilise ça sans le réapprendre.

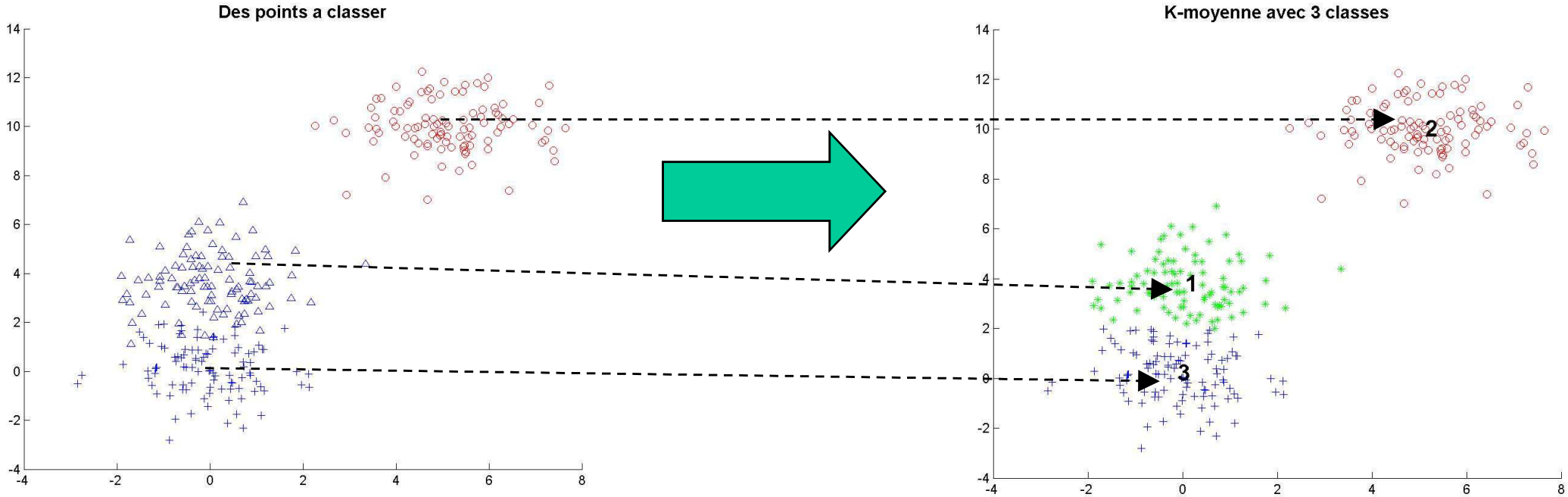
J'apprends ça (plus petit => moins d'exemples)

Quelques mots sur le non supervisé/estimation de features

Apprentissage non supervisé

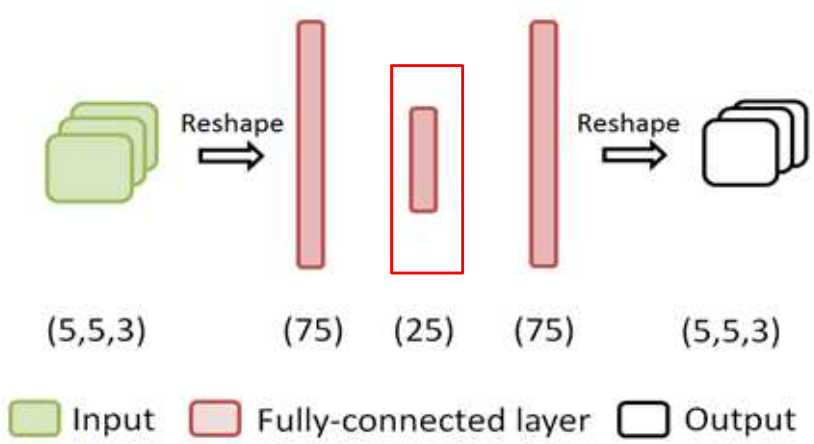
- Dans l'apprentissage **supervisé** : on dispose d'un ensemble de données labélisées à partir desquelles on cherche à déterminer un classifieur.
- Dans l'apprentissage **non supervisé :** on dispose d'un ensemble de données non labélisées que l'on désire regrouper. On cherche à comprendre leur structure.

Recherche de données « similaires »
suivant une métrique (euclidienne ici)

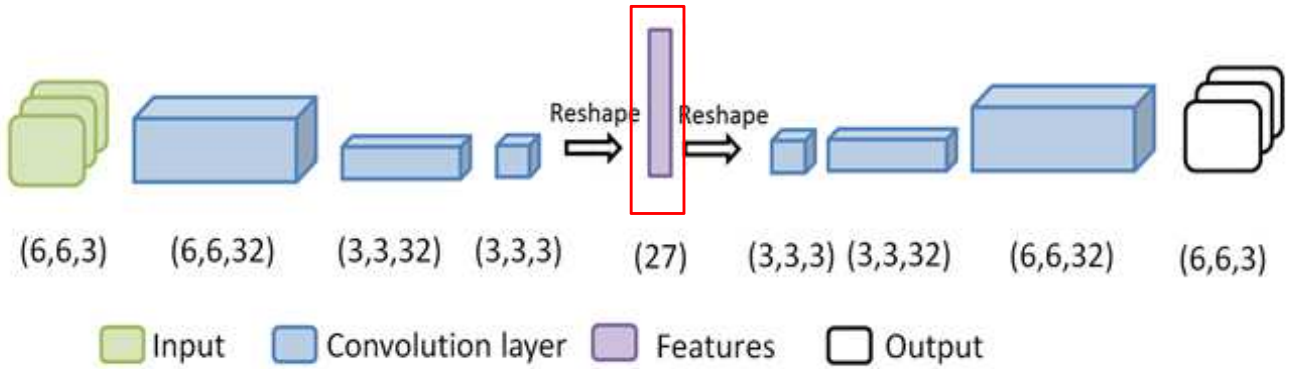


Apprentissage non supervisé

- Auto-Encoders



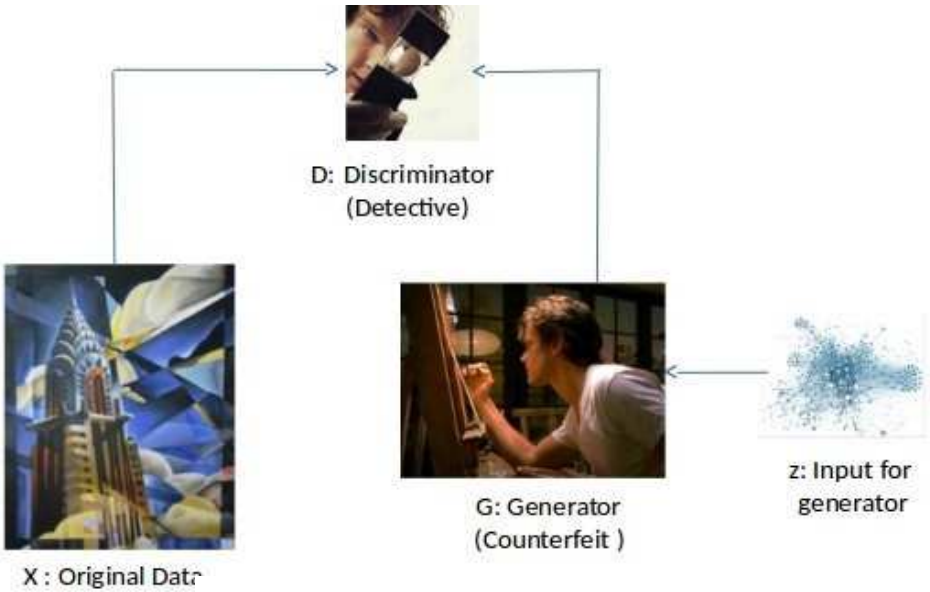
Dense Autoencoder

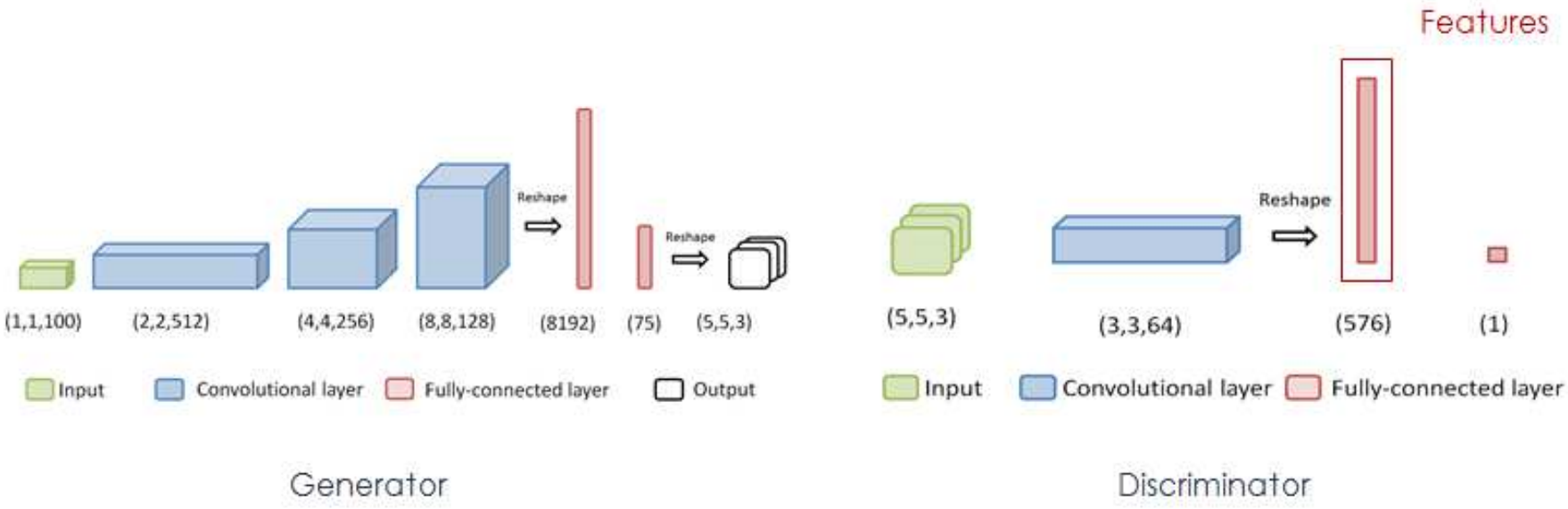


Convolutional Autoencoder

Apprentissage non supervisé

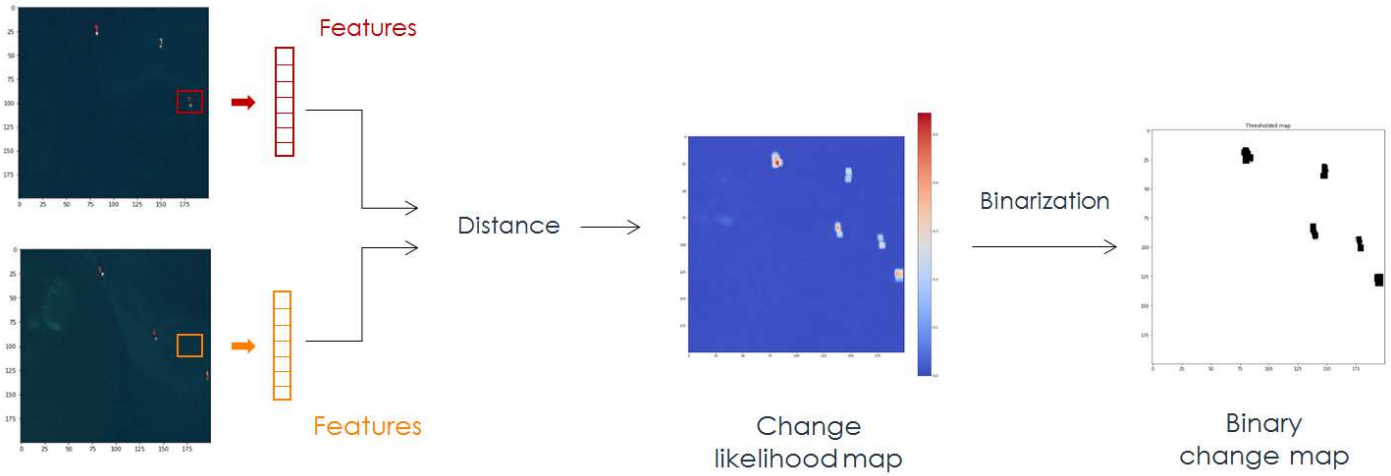
- Generative Adversarial Network (GAN)





Apprentissage non supervisé

- Example of features extraction for change detection



Apprentissage non supervisé

WIRED Google's Artificial Brain Learns to Find Cat Videos

BY WIRED UK 06.26.12 11:15 AM

The New York Times | http://nyti.ms/Lmw7zo

TECHNOLOGY

How Many Computers to Identify a Cat? 16,000

By JOHN MARKOFF JUNE 25, 2012

Mais qu'ont-ils donc fait ?

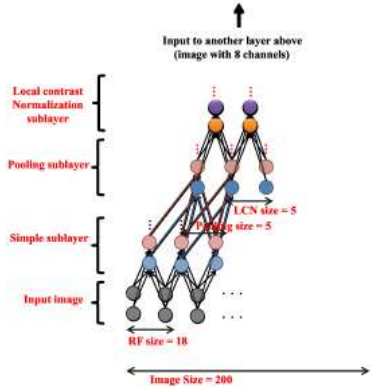
Apprentissage non supervisé

- **Le principe (Andrew Y. Ng, Ranzato & al, 2012)**

  - Création d'une base d'images non labélisées à partir de YouTube: 10 millions d' images 200x200 extraites de vidéos YouTube ® (1 image par vidéos pour éviter la redondance).
    NB: Grâce à un traitement de reconnaissance de visages, on sait que les visages représente moins de 3% des images, ce qui évite un surapprentissage des visages, on considère idem pour les chats.

10 millions d'images prises au hasard dans YouTube®

  - Choix d'un modèle d'apprentissage non supervisé: Ils font du « **Model engineering** »
    3 empilements de SAE+Pooling+Local Contrast Normalization.
    Utilisation de Local receptive fields (LeCUn 1998)
    limitant les connectivités d'une couche à l'autre,
    invariance par le L2 pooling+LCN.

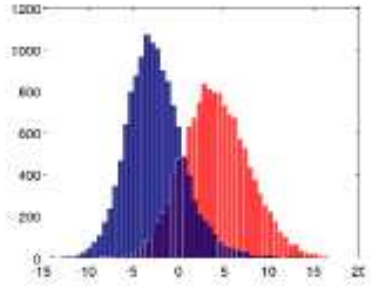


Modèle non supervisé: Un milliard de paramètres à estimer !
(NB: c'est encore 1 million de fois plus petit que le cortex visuel humain…)

Exécution sur 1000 machines (16000 cœurs) pendant 3 jours

Apprentissage non supervisé

- **Donc le modèle a « créé » un neurone qui permet de différencier les visages des non visages (faces/no faces) uniquement en lui présentant des données non labélisées.**



(Seuil= 0 ici)

Source: Ng, Ranzato&al 2012

Histograms of faces (red) vs. no faces (blue).

- **Les auteurs ont aussi calculé numériquement une image x* qui maximise la réponse de ce neurone :**

$$x^* = \arg\min_x f(x; W, H),$$

= sortie du neurone en test

Quelques buzz en cours (novembre 2018, non exhaustif)

Quelques buzz en cours

- Hybrid AI: le Lien entre « model-based « et « data-driven ».
- Robustesse aux « attaques »
- Weakly supervised
- Implémentation par « Separable convolution », Mobile net.
- Domain adaptation
- Deep attention network
- Deep reinforcement learning
- Etc.

Ou en est on aujourd’hui? (novembre 2018)

Ou en est on aujourd'hui (nov 2018)? Le Pour

- Le **Pour** des techniques data-driven basées sur le DL:
  - ➔Résultats performants sur de nombreux sujets, en un temps d'études relativement court
  - ➔Grande communauté, beaucoup d'outils didactiques, nombreuses bibliothèques en open-source => Les outils sont la (et en constante évolution) => on est en au stade du « model–engineering ».
  - ➔Pas besoin d'être un spécialiste du traitement d'images/Classif. pour utiliser le DL, bien que le choix du modèle peut-être parfois délicat.
  - ➔Nombreuses applications TI en Spatial: débruitage, pan-sharpening, ré-échantillonnage, détection de cibles, classification de scènes, etc.

- Ne pas oublier:
  - ➔Le DL et data-driven ne peuvet pas tout traiter => Les autres techniques restent d'actualité:
    - Data-driven: Random forest
    - Model-based, Hybrid Data driven-Model based
    - Etc.

Ou en est on aujourd'hui (nov 2018)? Le Contre

- Le **contre**
  - ➔Résultats empiriques. Pas de théorie mathématique (sauf optim.) : absence de borne théorique, etc. Cela implique un **problème de validation/explicabilité pour les systèmes critiques:** comment avoir une « **Good and Trustable Artificial Intelligence** ».
  - ➔Pas robuste aux attaques par génération de bruit adéquat dans une image paf exemple
  - ➔Si on veut faire du Deep Learning supervisé il faut de la puissance de calcul (cloud, etc.). Les systèmes actuels commencent à montrer une certaine limite.
  - ➔Reproductibilité des résultats: quand de grandes compagnies utilisent des gros moyens de calcul il est difficile pour les scientifiques de vérifier ces résultats
  - ➔Voir aussi transparent « DL Limites et Challenges »

Comment ne pas aller vers le 3ème hiver de l'IA?

- 2011: Watson (IBM) vs Jeopardy.
- 2011: Challenge Panneaux signalisation: Performances supérieures à l'humain
- 2012: Apprentissage non supervisé utilisant YouTube
- 2016: Alphago bat le champion du monde de Go.
- 2017: AlphaGo Zero bat AlphaGo par 100 jeux à zéro
- ***202x: 3ème hiver de l'IA ?***

Essayons d'éviter le 3ème hiver de l'IA: Problèmes à régler

- Limites et Challenge du DL
  - ➔« Small data. Une solution: le transfert learning.
  - ➔Coût: beaucoup de données, de paramètres à estimer, de paramètres à fixer: comment apprendre à apprendre?
  - ➔Améliorer encore les performances (hybrid AI?)
  - ➔Validation/certification (hybrid AI ?)
  - ➔Explicabilité/Interprétabilité
  - ➔Causalité
  - ➔Transparence, acceptabilité sociale
  - ➔Arriver à une « Trusatble good AI »

Merci pour votre attention